Image and Vision Computing 119 (2022) 104375

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Progressive ShallowNet for large scale dynamic and spontaneous facial behaviour analysis in children

Abdul Qayyum [a,e], Imran Razzak [b,*], Nour Moustafa [c], Moona Mazher [d]

[a] Department of Electrical and Computer Engineering, Dijon University, France
[b] School of Information Technology, Deakin University, Geelong, Australia
[c] School of Engineering and Information Technology, University of New South Wales, Canberra, Australia
[d] Department of Computer Engineering and Mathematics, University Rovira i Virgili, Tarragona, Spain
[e] ENIB, UMR CNRS 6285, LabSTICC, Brest, France

ARTICLE INFO

Article history:
Received 18 August 2021
Received in revised form 7 December 2021
Accepted 10 January 2022
Available online 17 January 2022

Keywords:
Psychological health
Human computer interaction
Emotion care
Depressed
Facial behavior recognition
Progressive ShallowNet
Patient monitoring

ABSTRACT

COVID-19 has severely disrupted every aspect of society and left negative impact on our life. Resisting the temptation in engaging face-to-face social connection is not as easy as we imagine. Breaking ties within social circle makes us lonely and isolated, that in turns increase the likelihood of depression related disease and even can leads to death by increasing the chance of heart disease. Not only adults, children's are equally impacted where the contribution of emotional competence to social competence has long term implications. Early identification skill for facial behaviour emotions, deficits, and expression may help to prevent the low social functioning. Deficits in young children's ability to differentiate human emotions can leads to social functioning impairment. However, the existing work focus on adult emotions recognition mostly and ignores emotion recognition in children. By considering the working of pyramidal cells in the cerebral cortex, in this paper, we present progressive lightweight shallow learning for the classification by efficiently utilizing the skip-connection for spontaneous facial behaviour recognition in children. Unlike earlier deep neural networks, we limit the alternative path for the gradient at the earlier part of the network by increase gradually with the depth of the network. Progressive ShallowNet is not only able to explore more feature space but also resolve the over-fitting issue for smaller data, due to limiting the residual path locally, making the network vulnerable to perturbations. We have conducted extensive experiments on benchmark facial behaviour analysis in children that showed significant performance gain comparatively.

1. Introduction

History tells us that society can be cohesive socially in crisis, and COVID-19 pandemic is presenting society with intimidating enemy that does not differentiate black and white. Resisting the temptation to engage face-to-face social connection is not as easy as we think. Breaking ties within our social circle makes us lonely, sad and isolated, that in turns increase the likelihood of depression related disease and even it can leads to death by increasing the chance of heart disease. People strive to cope with psychological dissonance in different ways such as adjusting them-self in new ways of living. Technology also play an important role in achieving their goal.

Not only adults, children's are equally impacted where the contribution of emotional competence to social competence has long term implications. Emotions are very essential for communication as well as for social interaction. Emotions play key role in decision making and are critical our daily life, i.e. how we engage with others and live our life, how we feel others emotions etc. No doubt, it is essential to remember, however, that no emotion is an island. Instead, we experience many emotions that are complex and nuanced, working simultaneously to make the varied and rich fabric of our emotional life. At times, it may seem that our emotions control us. The choices we make, the actions we take, and our perceptions are all affected by our emotions that we are experiencing at any given moment (See Fig. 1).

Emotions influence us, how we live and interact with computer. By knowing and understanding different types of human emotions, we can learn how emotions are expressed and how they can impact our behaviour. Recently, psychologists tried to understand the impact of emotions on our lives and identified several types of emotions. Several theories on emotions understanding have emerged and identify different categories and explain their behaviour that we feel. In 1970s, psychologist Paul Eckman, categorized emotions into six basic categories:

[*] Corresponding author.
E-mail address: imran.razzak@deakin.edu.au (I. Razzak).

https://doi.org/10.1016/j.imavis.2022.104375

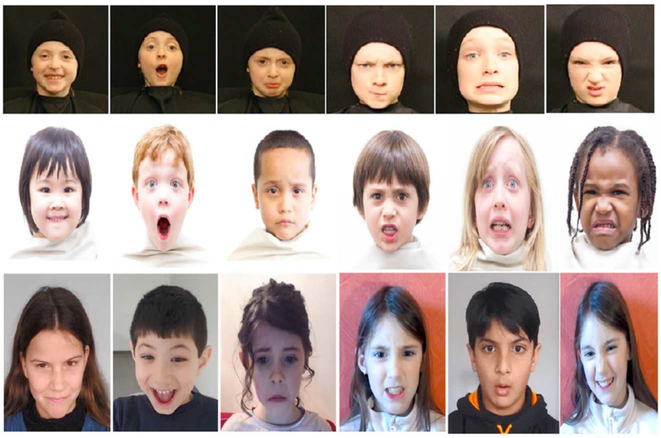


Fig. 1. Spontaneous facial expressions.

disgust, fear, happiness, sadness, surprise, and anger. These emotions are experienced all over the world. Later on, Paul extended the list of emotions to 10 emotions: embarrassment, shame, pride and excitement. In the adult, even in early life such as preschool, the contributions of emotion to social competence have long-term implications on our life.

Facial expressions are one of the efficient and effective means to convey the information on one's emotional feelings to the others. Analysis of facial expression is very important for designing an efficient and learnable user interface for effective communication. Deficits in the ability to understand one's facial expressions may lead us to the impairments in social functioning. Denham et al. showed that academic achievements and peer-rated popularity are strongly correlated with understanding the others' emotions [5,6]. Development of emotion in human is the largest and productive research areas in psychology. For decades, researchers have been fascinated by it and exploring how we respond, identify, and interpret the other's expression. Most of the research on emotional behaviour identification has relied on controlled stimulus sets of adults emotions. However, there is less work on understanding children emotions. Recently, the stimulus set of emotional facial expressions (Child Affective Facial Expression, CAFE) is introduced consisting of ethnically and racially diverse childern with ages between 2 and 8-year old posing basic six basic expressions-fearful, angry, happy, disgusted sad, and surprised, as well as neutral face (See Fig. 2).

Children express their emotions through six expressions that are surprise, fear, anger, disgust, anger, sadness, and happiness. Understanding children emotions at early stage could help to deal with poor social functioning though efficient user interface. Most of the existing efforts focus on facial expression recognition in adults with posed expression (artificially generation emotions). However, understanding the behaviour in children especially the case of spontaneous facial expression is considered less comparatively. Beside, analyzing spontaneous expressions may help to deal with deficit through development of efficient user interface. To deal aforementioned challenge, in this work, we explored the problem of spontaneous emotion analysis in children and proposed a novel end-to-end framework consisting of lightweight ShallowNet. The key contributions of this work can be described as

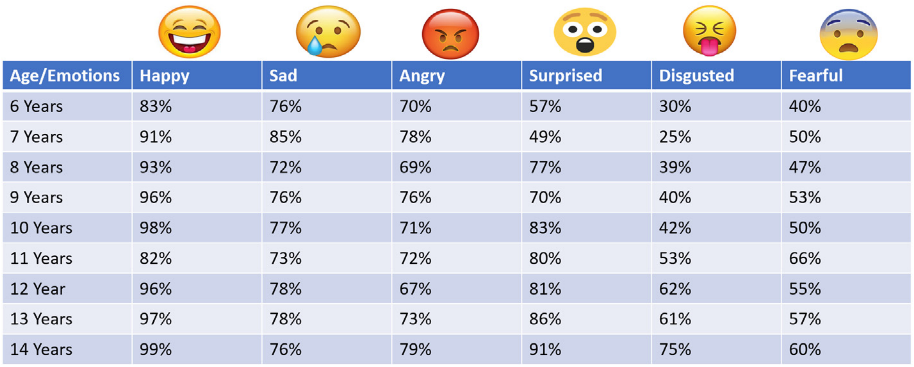
- To investigate the underlying principles of Child–Computer interaction, We present novel progressive weight ShallowNet for the analysis of spontaneous facial emotions in children.
- Unlike earlier networks, we limit the alternative path for the gradient at the earlier stage of the network and increase gradually with the depth of network.
- Progressive light ShallowNet is able to explore more feature space and vulnerable to perturbations, thus it helps to overcome over fitting challenge due to lower number of residual connections.
- Extensive experiment are conducted on benchmark facial behaviour analysis in children for investigation of child behaviour during interaction with computer that showed significant improvement in detection performance comparatively.

The rest of the paper is organized as: in the next section, we present the state of the art facial expression recognition and recent development on spontaneous facial expression. In Section 3, we present an alternative path based deep learning network followed by progressive light ShallowNet and its application on child spontaneous behaviour identification of child based on facial expression. In Section 5, we present experimental results in detail.

2. Related work

To improve emotion recognition performance, recently, different deep networks are individually or sequentially in different hierarchically. Dense neural networks are trained to identify the areas related to facial expression followed by classification through autoencoder [15]. Similarly, Liu et al. presented boasted dense neural network that perform feature representation, feature extraction and classification iteratively [14] which results in propagating the classification error to initiate the feature selection until it converges. To find local-translation-invariant representation, Rifai et al. presented multi-scale contractive neural network that identify the emotion related factors from pose and identify hierarchically [17].

Recognition of spontaneous facial expression in children is complex and challenging comparatively. Unlike pose expression, there are large variations and poses, whereas spontaneous expression has lot of real world applications such as daily interactions, meetings, and debate



Age/Emotions	Happy	Sad	Angry	Surprised	Disgusted	Fearful
6 Years	83%	76%	70%	57%	30%	40%
7 Years	91%	85%	78%	49%	25%	50%
8 Years	93%	72%	69%	77%	39%	47%
9 Years	96%	76%	76%	70%	40%	53%
10 Years	98%	77%	71%	83%	42%	50%
11 Years	82%	73%	72%	80%	53%	66%
12 Year	96%	78%	67%	81%	62%	55%
13 Years	97%	78%	73%	86%	61%	57%
14 Years	99%	76%	79%	91%	75%	60%

Fig. 2. Learning of facial expressions in children

summarization. Recently, frontal face images are localized to preserve the facial expression using generative adversarial networks [13], followed by facial expression recognition using discriminator. Similarly, GAN is used to generate images with different expression under arbitrary poses [22]. Chen et al. combined VAE and GAN to preserve the representation, which is able to disentangled the identity explicitly and generate facial synthesis while preserving the images [21].

Sequential methods such as recurrent neural networks have been applied model the semantic relationship of different face points. Chanti et al. applied grid convolution to encoder the spatial correlation and used LSTM to model the temporal relationship [2]. In another work, Tran el. [19] all used 3D convolution with shared weights [1,3,4,7,23]. followed by LSTM. Similarly, Vielzeuf et al used 3D convolotional kernal to compute structural score from each consective frame [20]. To model the time varying spontaneous expression, casecaded CNN and LSTM has also been used to model the sequential data [7,12].

Facial expression recognition has made substantial progress, however, most of the work focus on posed expression whereas real world scenarios is spontaneous expression. Beside, efforts for spontaneous expression recognition in children is comparatively less considered, whereas early detection of emotion deficits may help to prevent social functioning in later age.

3. Child–computer interaction facial behaviour analysis

Even though decades of research on understanding emotions role in user interface, there is a little understanding of how it directly manipulate our emotional compliment, undermines, or orthogonal to aspects of interface that specifically addresses the user's need. The emotional aspects play critical role in inclusive system based on integrating the persons with disabilities. Recently, identification of human emotions have made significant progress in the past few years. However, children's spontaneous facial expression recognition has received considerably less attention, especially that pose-invariant occlusion-robust children's facial expressions in real-world scenarios have received significantly less attention. The deficit in early understanding of emotion as well as in skills of expression results in poor social functioning. Designing interface should consider application specific. Interface that fails to manifest or ignores user's emotions can significantly impact the performance and risks being perceived as socially inept, untrustworthy, and incompetent. To deal with aforementioned challenges, we analyzed the spontaneous children facial expression and present lightweight progressive ShallowNet. The proposed framework is able to extract better features due to less number of residual connection locally, hence the proposed network is vulnerable to perturbations, which also solve the overfitting issue for small data. Unlike the existing ResNet and its variants, we reduced the residual connections at earlier phase of network and graduate increased with the depth of the network. In the below section, we first discussed the proposed lightweight shallow network, followed by its applicability on the spontaneous facial expression. Fig. 3 illustrate the proposed lightweight ShallowNet.

3.1. Progressive light weight ShallowNet

Most the existing deep network for understanding emotions are over parameterized. The network that has hidden units less number of polynomially results in better classification performance comparatively than over-parameterized deeper networks. For example, each layer in DenseNet takes additional input from its preceding layers and passes it to all its following layers. Thus, each layer in DenseNet is receiving a “collective knowledge” from all its preceding layers. However, it does not only result in very complex network but also over patrametrized. Similarly, Residual neural network is also over parameterize. Thus, we can say, it is natural to have numerous global optima and large training loss. In this study, we propose progressive light weight Shallow network (ShallowNet) with alternative shortcut path by leveraging the pyramidal cells in the cerebral cortex learning. Unlike, earlier networks, we decreased the number of skip connection at early stage, forcing the network to learn better representation. The proposed progressive ShallowNet architecture with additional transition block at each stage is shown in Fig. 4. Notice that the first stage in proposed ShallowNet consists of only two residual units with identify function and transition block which acts as a projection shortcut. This enables the ShallowNet to consider only features with semantic information by limiting the jump over connections. Notice, that earlier part of the network block is smaller in size, however, with the depth of network, we have increased the size of blocks that force to extract basic features with semantic information in first part of network and focus on high-level features at second part of the network. Beside this, it also speedup the learning by reducing the impact of vanishing gradient, as only few of the layers are propagated through each stage. Moreover, it allows the network to consider shallow features consisting of semantic information but not deep at each stage progressively, which helps prevent over-fitting for small data and capture feature with different level of detail. As the network grows, it gradually restores the skipped layers.

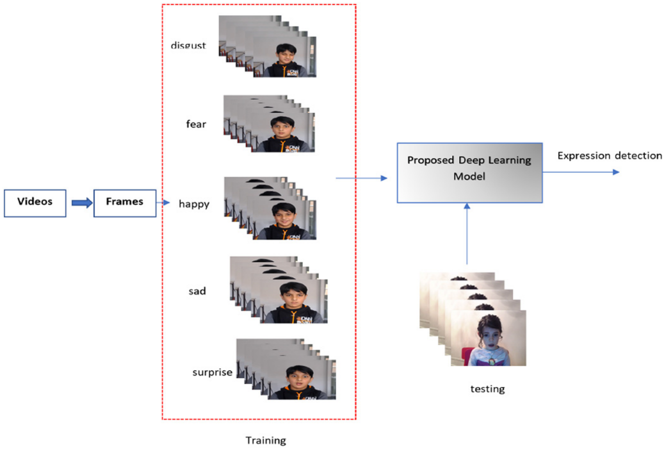


Fig. 3. Proposed framework for spontaneous facial expression recognition.

Notice (Fig. 4) that we have divided the framework into several smaller blocks. Each block further have of different number of residual unit which are different in structure (residual unit). Fig. 4 shows the proposed framework consisting of block approach. We can observe that earlier stages of network consist of smaller size residual unit with less number of skip connections. The output of each stage is forward to the transition layer, which acts as the projection matrix.

The progressive lightweight ShallowNet is based on working of pyramidal cells in the cerebral cortex. Similar to pyramidal cells, we develop small unit size in earlier blocks with less alternative path to carry additional low-level information, which helps to jump over the layers within each unit and maintain the dense connectivity. Thus, the resultant ShallowNet is less complicated, carry useful feature information and make it possible to learn hundreds of layers with much better accuracy and comparatively less training loss. Fig. 4 illustrates the proposed progressive light network architecture with alternative skip connections. Notice that we have also used the supervision layer to improve the performance, resulting in propagation of the error signal to earlier blocks efficiently and directly. This addition of implicit in-depth supervision layer in earlier small blocks with alternative paths can direct the supervision from the final block.

Traditionally, the feed-forward convolutional neural network connects the output of the mth layer as the input to $(n+1)$th layer which gives the rise to its following transition layer $x_m = H_n(x_n 1)$. Residual neural network consist of alternative path that uses identify function to bypasses the non-linear transformations

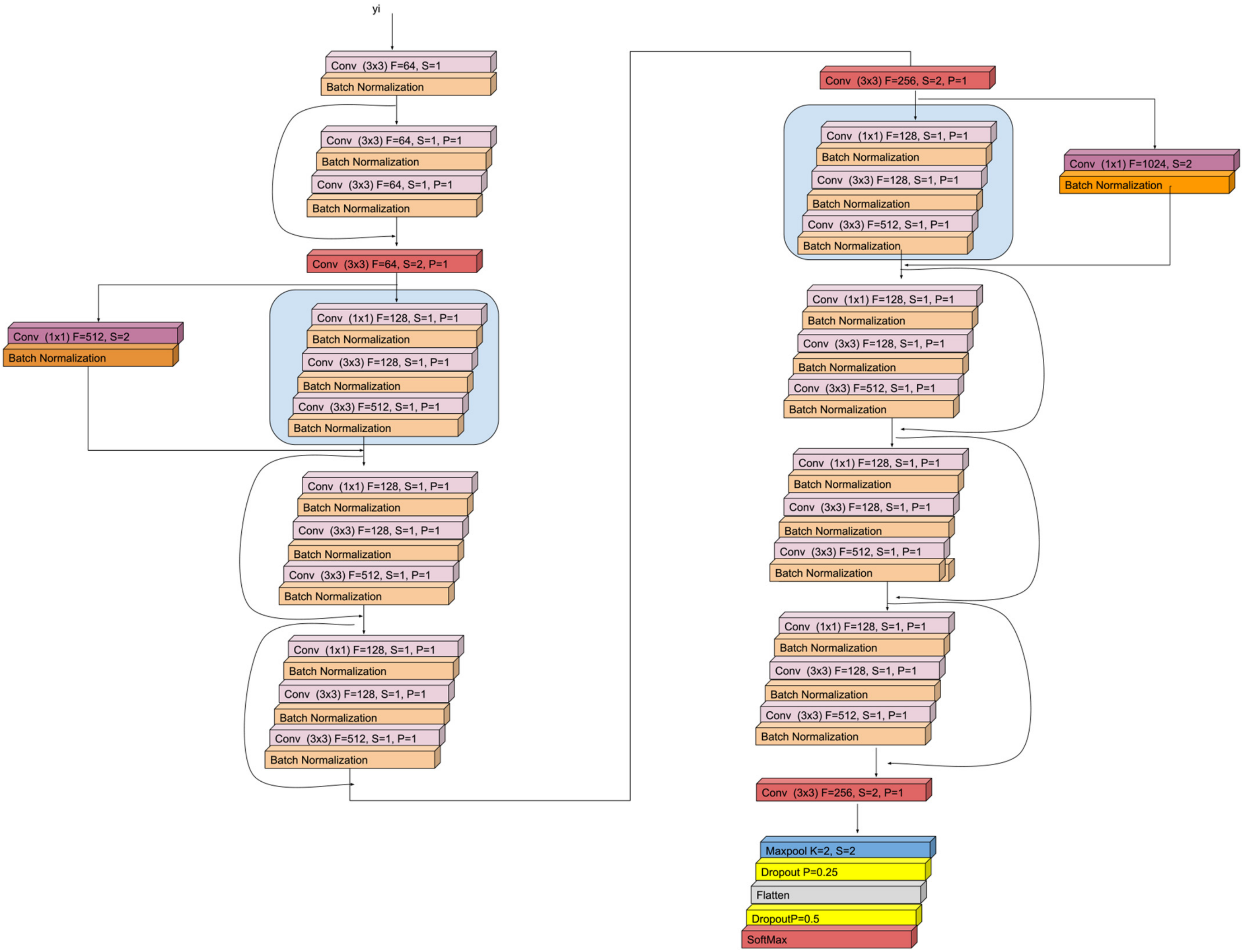

yi
Conv (3x3) F=64, S=1
Batch Normalization
Conv (3x3) F=64, S=1, P=1
Batch Normalization
Conv (3x3) F=64, S=1, P=1
Batch Normalization
Conv (3x3) F=64, S=2, P=1
Conv (1x1) F=512, S=2
Batch Normalization
Conv (1x1) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=512, S=1, P=1
Batch Normalization
Conv (1x1) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=512, S=1, P=1
Batch Normalization
Conv (1x1) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=512, S=1, P=1
Batch Normalization
Conv (3x3) F=256, S=2, P=1
Conv (1x1) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=512, S=1, P=1
Batch Normalization
Conv (1x1) F=1024, S=2
Batch Normalization
Conv (1x1) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=512, S=1, P=1
Batch Normalization
Conv (1x1) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=512, S=1, P=1
Batch Normalization
Conv (1x1) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=128, S=1, P=1
Batch Normalization
Conv (3x3) F=512, S=1, P=1
Batch Normalization
Conv (3x3) F=256, S=2, P=1
Maxpool K=2, S=2
Dropout P=0.25
Flatten
DropoutP=0.5
SoftMax

Fig. 4. Architectures of proposed three stage progressive lightweight ShallowNet.

$$x_n = H_n(x_{n1}) + x_n - 1.$$

In our case, the proposed progressive light weight ShallowNet network further improves the flow between layer thorough alternative path within each block. Thus, we can write

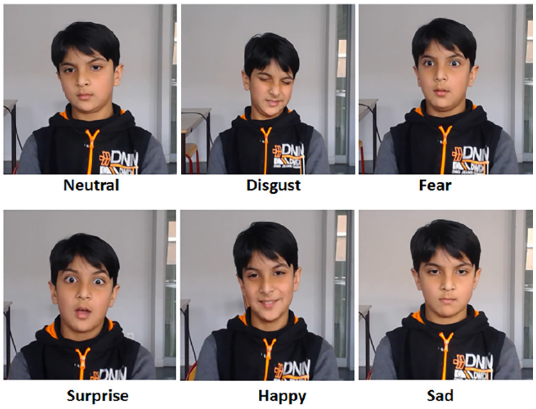


Fig. 5. Six common facial emotion.

$$U_n = H_n([U_1, U_2, \ldots, U_i] + x_{n-1}).$$

Unlike Residual neural network, We have efficiently utilized the alternative path by dividing each units into multiple subparts and adopted the alternative path in each block and used the block's projection path to block change, thus our proposed ShallowNet results in learning the collective features from all earlier layers. In addition to this, ShallowNet is able to go deeper than its counterpart. Each network block in ShallowNet consists of multiple skip path; however, the size of unit grows with the depth of the network, i.e. the initial block consist of 2 convolutional layers with one unit, where as next block consists of three convolutional layers with three units. Thus, the proposed network has less channels comparatively, making it thinner and deeper and extracting best feature representation (See Fig. 5).

Table 1
ShallowNet parameter.

Parameters	Values
Epochs	25
Initial learn rate	1×10^{-3}
Validation frequency	300
Momentum	0.95

Table 2
Evaluation of proposed lightweight ShallowNet (%).

	Support	Recall	Precision	Recall	F1-score
Disgust	97	1.00	1.00	1.00	1.00
Fear	529	0.98	0.99	0.98	0.99
Happy	778	0.99	1.00	0.99	0.99
Sad	512	1.00	0.99	1.00	0.99
Surprise	546	0.99	0.98	0.99	0.98
Macro avg.	2462	0.99	0.99	0.99	0.99
Weighted avg.	246	**0.993**	**0.994**	**0.993**	**0.992**

The traditional network with L number of layers consist of L connections-one between each layer, however, the ShallowNet has $D \times \frac{(U+1)U}{2}$ connections in each layer whereas U is the number of units and D is the depth of the network om each block. In each layer of residual network, several parameters are directly proportional to $C \times C$, whereas in ShallowNet the parameters are directly proportional to $U \times K \times K$. As $C >>> U$ in ShallowNet thus, we can say that the proposed ShallowNet has much smaller parameters comparatively. Fig. 4 illustrate the key block of ShallowNet. Notice that the size of feature map sizes is the same within the block.

The identity shortcuts for same dimensions is written as

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{1}$$

3.2. Child–spontaneous interaction

In our early age, we learn to produce and differentiate different facial expression and play key role in decision making and are critical our daily life. Recent studies shown that children with emotion recognition are better not only studies but also in social interactions [9,10,16]. Understanding emotions are linked various outcomes and are critical for cognitive development [18]. Early deficit in emotion and expression identification may help to deal with poor social functioning. Generally, facial expression is spontaneous, thus in this work, we present an end to end recognition of spontaneous facial expression in children. We present light weight shallow-net by limiting residual connections as shown in Fig. 3.

3.3. Dataset

Recently, Khan et al. presented children's spontaneous emotional database (LIRIS-CSE) consisting of children from diverse ethnicities. Video clips/dynamic images were recorded in diverse setting with six core emotions such as “happiness”, “sadness”, “disgust”, “anger”, “surprise”, and “fear”. The expression are same as of other facial expression datasets for adults. The unconstrained videos are recorded with the constraint-free environment such as children were free to move their head and hand in the free sitting environment. There were no restriction on kids were applied during dataset collection while they were watching special build images. The dataset was labelled and rated by 22 human rates. LIRIS-CSE consists of 26,000 frames of kids emotions. As kids were recorded in constraints free environment, thus, their expression is natural and spontaneous.

Table 3
Comparative evaluation of proposed framework with its counter network.

Model	Precision	Recall	F1-score
Proposed Shallownet network	**0.994**	**0.993**	**0.992**
DenseNet-freeze	0.94	0.95	0.94
ResNet-freeze	0.97	0.97	0.97
Inception-freeze	0.97	0.97	0.97
MobileNet-freeze	0.95	0.95	0.95
DenseNet-fine-tuned	0.88	0.88	0.88
ResNet-fine-tuned	0.90	0.89	0.89
Inception-fine-tuned	0.90	0.90	0.90
MobileNet-fine-tuned	0.88	0.87	0.87
Khan et al. [11]	0.81	0.82	0.83

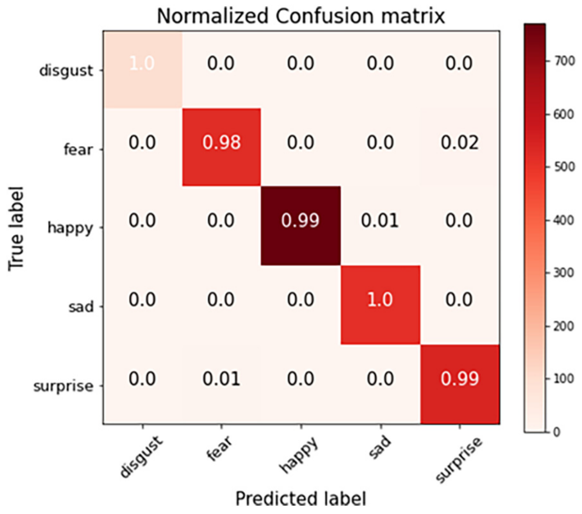


Fig. 6. Confusion matrix.

4. Experiment

User Interface design should mainly focus on application specific goals. Even though years of research, there is a little understanding that how a user interact is directly impacted by our emotion. Recently, identification of human emotions have made significant progress in the past few years. However, children's spontaneous facial expression recognition has received considerably less attention, especially that pose-invariant occlusion-robust children's facial expressions in real-world scenarios have received significantly less attention. Children's spontaneous facial expression recognition has received considerably less attention, especially that pose-invariant occlusion-robust children's facial expressions in real-world scenarios have received significantly less attention. The deficit in early understanding of emotion and expression skills results in poor social functioning. In this section, we present experimental setup, results and evaluation of proposed framework on benchmark child spontaneous facial expression dataset. Tables 4 and 5

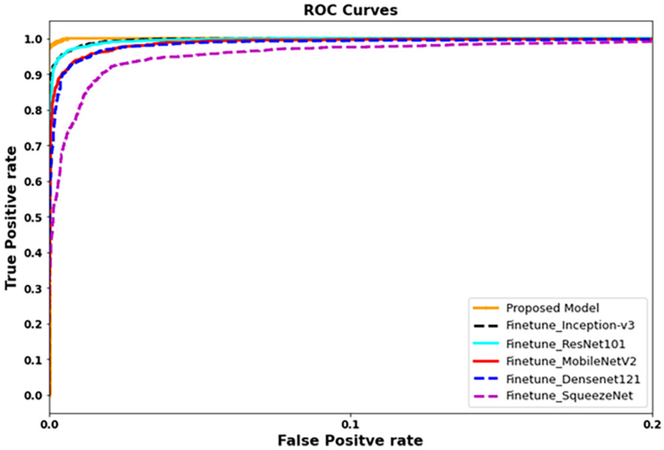


Fig. 7. Comparison of ROC curve of proposed progressive light residual learning with state of the art transfer learning methods using freeze scenario.

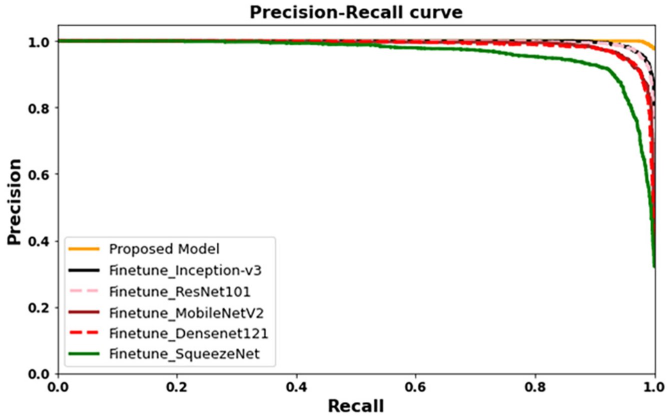


Fig. 8. Comparison of Precision–Recall curve of proposed progressive light residual learning with state of the art transfer learning methods using freeze scenario.

and Figs. 7 and 8 illustrate the experimental results. To generalize the performance gain, we have performed 10-fold cross-validation and compared the performance using different evaluation measures such as precision, recall, F-score, specificity, sensitivity, specificity and area under the curve (AUC).

4.1. Network parameter

Progressive lightweight ShallowNet architecture consists of smaller initial units with additional supervision block and transition blocks. Initial blocks consist of 2 convolution layer of 3 × 3 followed by batch normalization. Block 1 is followed by an alternative project path. Block 2 consists of 2 subunit, consisting of convolution layers (3 × 3 and batch normalization followed by transition and supervision blocks). To learn the best possible feature representation, ShallowNet has fewer alternative paths with smaller unit size in earlier blocks. We have increased the number of alternative path and unit size in later stages. We have initialized the weights as in [8] and trained the network from scratch. In order to find the best parameters and see the impact of the parameter in each block, we have trained ShallowNet network on different parameter values for each block. We have set the learning rate to (1×10^{-3}) and momentum of (0.95). We have stopped the training when there is no change in error rate for 25 epochs. Notice that the size of the kernel is continuously increasing with the growth of the network. Table 1 describes the training parameters.

4.2. Results

To visualize the robustness of network, we performed different experiments with different parameters and variations in network structure i.e. 2 Blocks, 3 Blocks with variable residual units. In order to generalize the performance, we have performed k-fold cross-validation ($k = 10$) and compared the performance with start-of-the-art methods. We have evaluated the performance using different evaluation metrics such as the precision, recall, F-score, sensitivity, specificity and area under the curve (AUC). Table 4 describes the emotion recognition results for each class. Notice that, we have achieved high accuracy for disgust emotions followed by ‘happy’. Similarly, we can notice the trend for other emotions. Fig. 6 shows the confusion matrix of the proposed progressive light residual network. It shows that proposed ShalloweNet achieved significantly better performance comparatively, however, it is worthy to highlight that performance is slightly poor for ‘fear’ (See Tables 2 and 3).

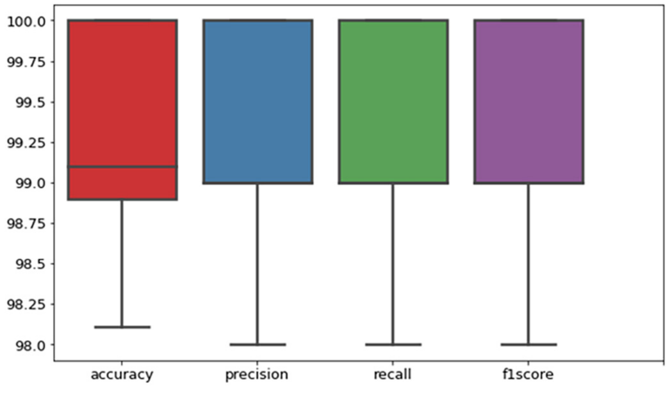


Fig. 9. Analysis of performance of proposed progressive light weight ShallowNet.

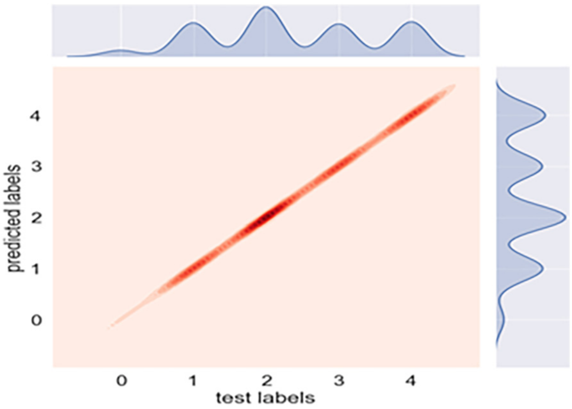


Fig. 10. Comparison of predicted vs test values.

Tables 4 and 5 showed that progressive lightweight ShallowNet network achieve best performance 0.99 (F1 score), 0.99 (recall) and 0.99 (precision) on 2462 images. Notice that disgust and happy expressions showed better performance with almost no false rate (see confusion matrix in Fig. 6). In addition to this, we have also applied transfer learning using several state-of-the-art methods, especially counter networks such as Inception, MobileNet, ResNet, and DenseNet. We have considered both freeze and fine turning in transfer learning. We applied the transferability of deep models trained for facial expression recognition on adults expression dataset (source) to the children expression (target). Tables 5 and 4 shows that proposed progressive lightweight ShallowNet showed significant improvement results compared to the baseline methods as as as state of the art transfer learning methods.

4.3. Discussion

Unlike, earlier work on adult or child facial recognition based on posed expression (fake or disguise inner feeling), we focused on solving the problem of expression in spontaneous expressions captured in an unconstrained environment. In this section, we have compared the performance with state of the art methods (See Figs. 9 and 10).

Table 4
Comparative analysis (average) of ShallowNet framework with its counter network (Inception, MobileNet, DenseNet, ResNet and SqueezNet).

	Average Accu	Average precision	Average recall	Average F1-score
Proposed	**99.06**	**99.16**	**99.22**	**99.19**
SqueezNet	86.92	89.25	82.20	84.46
DensNet121	87.65	86.95	86.94	86.78
ResNet101	90.37	89.98	84.28	86.41
Inception_V3	89.27	90.26	88.83	89.45
MobileNet-v2	87.32	85.40	88.68	86.71
Deep-CNN [11]	77.23	69.43	77.88	81.44

Table 5
Comparative analysis (average) of proposed framework with its counter network (ResNet, MobileNet, DenseNet and SqueezNet using Freeze based fine tuning).

	Accuracy	Precision	Recall	F1-score
Proposed	**99.06**	**99.16**	**99.22**	**99.19**
SqueezNet	86.92	89.25	82.20	84.46
DensNet121	87.65	86.95	86.94	86.78
ResNet101	90.37	89.98	84.28	86.41
Inception_V3	89.27	90.26	88.83	89.45
MobileNet-v2	87.32	85.40	88.68	86.71
Deep-CNN [11]	77.23	69.43	77.88	81.44

In addition to this, we have applied transfer learning using several benchmark methods, especially counter networks such as ResNet, MobileNet, Inception and DenseNet. We have considered both freeze and fine turning in transfer learning. Tables 4 and 5 shows that proposed progressive lightweight ShallowNet showed significant improvement results compared to the baseline methods as as as state of the art transfer learning methods. We can notice that our proposed network showed the significantly better performance to 99.18% in comparison to 97.01% and 96.81% for Inception and ResNet, respectively, as our approach is a variant of ResNet. The significant gain in classification performance compared to Resnet showed the robustness of the proposed network for complex and relativity small dataset problems. From Tables 5, 4 and Figs. 7, 11, and 8 we can notice that fine-tuning showed considerably better performance in comparison to freeze network. Furthermore, we can also notice that our proposed approach achieved significantly better performance than state-of-the-art methods (See Fig. 12).

Comparing with benchmark emotion classification methods, we have the following interesting observations.

- Progressive light weight ShallowNet converges faster compared to counter network.
- Progressive lightweight ShallowNet can extract better representation due to fewer skip connection at the earlier part of the network.
- Progressive ShallowNet showed better performance with the increase of network depth for the larger dataset.
- Progressive light ShallowNet has much lesser parameters comparatively.
- The proposed ShallowNet showed a significant gain in performance and reduced the top error successfully. This shows the effectiveness of proposed progressive light network over its counter network, which validates its robustness, hence it can be applied to many other classification problems efficiently.

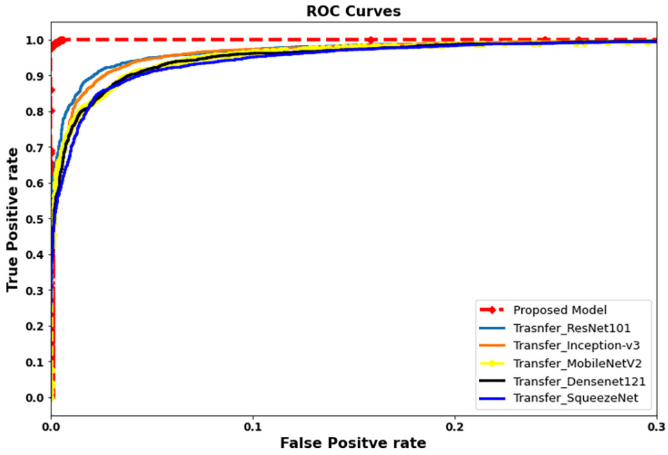


Fig. 11. ROC curve comparative analysis with transfer learning based on fine-tuned network.

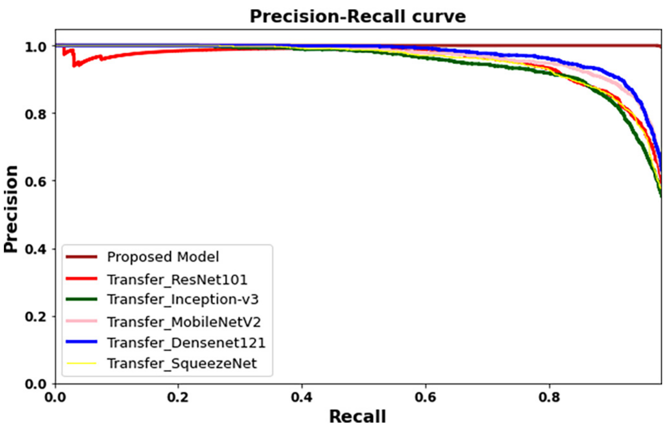


Fig. 12. ROC curve comparative analysis with transfer learning based on fine-tuned network.

5. Conclusion

In this paper, we addressed the role of children's emotions in human computer interaction. Emotions are an important part of the human–computer interaction especially in children. Multimodal interfaces that include facial expression and voices can manifest large range of nuanced emotions than was possible in purely textual interfaces. Interface that fails to manifest or ignores user's emotions can significantly impact the performance. We presented light weight progressive ShallowNet learning to classify spontaneous emotion recognition in children. Unlike earlier residual network, we limit the residual connection in the earlier phase of network and increased gradually with the growth of network which can learn robust features due to less number of residual connection. Hence, the proposed network is vulnerable to perturbations which helps to overcome overfitting issue for smaller data. Extensive experimental on benchmark children facial expression dataset showed that proposed framework is significantly better with performance to 99.16% in comparison to 90.37%, 86.82, and 87.65% using ResNet, DenseNet and SequezeNet, respectively. The significant gain in performance compared to its counter network showed the robustness of proposed ShallowNet for complex and relativity small dataset problems.

Credit author statement

Abdul Qayyum, Mona Mazhar and Imran Razzak performed experiment. Mona Mazhar and Nour Mustafa analyzed the results and Imran Razzak and Nour Moustafa wrote the manuscript.

Declaration of Competing Interest

Authors declare no conflict of interest.

References

[1] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, S. Lucey, Using synthetic data to improve facial expression analysis with 3d convolutional networks, Proceedings of the IEEE International Conference on Computer Vision Workshops 2017, pp. 1609–1618.

[2] D.A. Al Chanti, A. Caplier, Deep learning for spatio-temporal modeling of dynamic spontaneous emotions, IEEE Trans. Affect. Comput. 12 (2) (2021) 363–376, https://doi.org/10.1109/TAFFC.2018.2873600.

[3] P. Barros, S. Wermter, Developing crossmodal expression recognition based on a deep neural model, Adapt. Behav. 24 (5) (2016) 373–396.

[4] Spatio-Temporal, Samples using deep. Audio-visual emotion recognition system for variable length spatio-temporal samples using deep transfer-learning. Business Information Systems: 23rd International Conference, BIS 2020, CO, USA June 8-10, 2020, Proceedings, volume 389Colorado Springs, Springer 2020, p. 434.

[5] S.A. Denham, E.A. Couchoud, Young preschoolers' understanding of emotions, Child Stud. J. 20 (3) (1990) 171–192.

[6] S.A. Denham, M. McKinley, E.A. Couchoud, R. Holt, Emotional and behavioral predictors of preschool peer ratings, Child Dev. 61 (4) (1990) 1145–1152.

[7] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using cnn-rnn and c3d hybrid networks, Proceedings of the 18th ACM International Conference on Multimodal Interaction 2016, pp. 445–450.

[8] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 1026–1034.

[9] S.M. Jones, J.L. Brown, W.L. Hoglund, J.L. Aber, A school-randomized clinical trial of an integrated social-emotional learning and literacy intervention: Impacts after 1 school year, J. Consult. Clin. Psychol. 78 (6) (2010) 829.

[10] S.M. Jones, J.L. Brown, J.L. Aber, Two-year impacts of a universal school-based social-emotional and literacy intervention: an experiment in translational developmental research, Child Dev. 82 (2) (2011) 533–554.

[11] R.A. Khan, A. Crenn, A. Meyer, S. Bouakaz, A novel database of children's spontaneous facial expressions (LIRIS-CSE), Image Vision Comput. 83 (2019) 61–69.

[12] D.H. Kim, W.J. Baddar, J. Jang, Y.M. Ro, Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition, IEEE Trans. Affect. Comput. 10 (2) (2017) 223–236.

[13] Y.-H. Lai, S.-H. Lai, Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition, 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE 2018, pp. 263–270.

[14] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition 2014, pp. 1805–1812.

[15] Y. Lv, Z. Feng, C. Xu, Facial expression recognition via deep learning, 2014 International Conference on Smart Computing, IEEE 2014, pp. 303–308.

[16] D.J. McDowell, R. O'Neil, R.D. Parke, Display Rule Application In A Disappointing Situation and Children's Emotional Reactivity: Relations With Social Competence. Merrill-Palmer Quarterly (1982-), American Psychological Association, USA, 2000 306–324.

[17] S. Rifai, Y. Bengio, A. Courville, P Vincent, M. Mirza, Disentangling factors of variation for facial expression recognition, European Conference on Computer Vision, Springer 2012, pp. 808–822.

[18] M. Sprung, H.M. Münch, P.L. Harris, C. Ebesutani, S.G. Hofmann, Children's emotion understanding: a meta-analysis of training studies, Dev. Rev. 37 (2015) 41–65.

[19] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 4489–4497.

[20] V. Vielzeuf, S. Pateux, F. Jurie, Temporal multimodal fusion for video emotion classification in the wild, Proceedings of the 19th ACM International Conference on Multimodal Interaction 2017, pp. 569–576.

[21] H. Yang, Z. Zhang, L. Yin, Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks, 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE 2018, pp. 294–301.

[22] F. Zhang, T. Zhang, Q. Mao, C. Xu, Joint pose and expression modeling for facial expression recognition, Proceedings of the IEEE conference on computer vision and pattern recognition 2018, pp. 3359–3368.

[23] J. Zhao, X. Mao, J. Zhang, Learning deep facial expression features from image and optical flow sequences using 3d cnn, Visual Comput. 34 (10) (2018) 1461–1475.